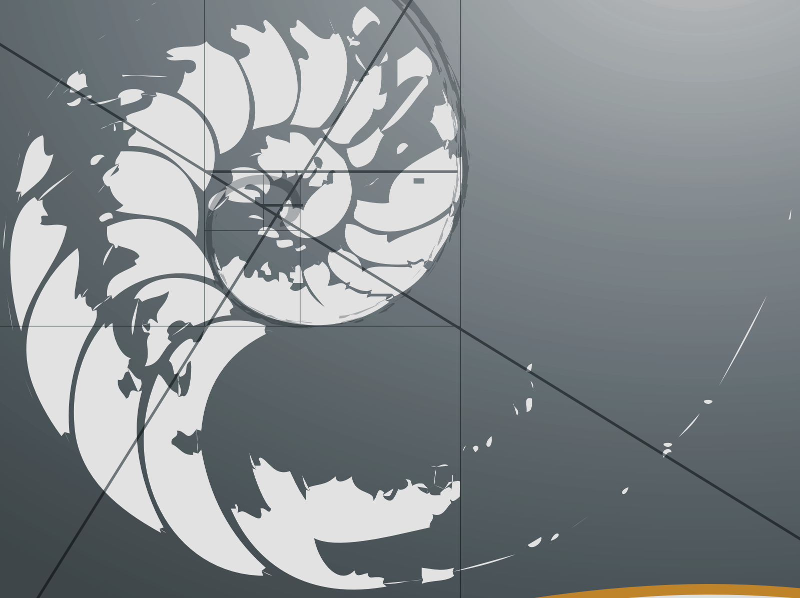


National Science Foundation
Advisory Committee for Cyberinfrastructure
Task Force on Data and Visualization

Final Report, March 2011

Data and Visualization Task Force

ACCI Members

Atkins, Dan
University of Michigan

Baker, Shenda
Harvey Mudd College

Dietterich, Thomas
Oregon State University

Feldman, Stuart
Google

Hey, Tony
Microsoft Research

Lyon, Liz
UKOLN

Community Members

Almes, Guy
Altinas, Ilkay
Burt, Ruth
Corbato, Steven
Denisov, Dimitri
French, Jim
King, Gary
McGee, John
Messina, Paul
Pellegrino, Don
Stoddard, Victoria
Thorley, Mark
Wilson, Bruce
Wu, Cathy
Hirsh, Haym
Papka, Mike
Andrews, Phil

National Science Foundation Liaisons

Blatecky, Alan
Bogden, Philip
Friedlander, Amy
Meacham, Stephen
Pennington, Rob

Note: Although this report was prepared by a task force commissioned by the National Science Foundation, all opinions, findings, and recommendations expressed within it are those of the task force and do not necessarily reflect the views of the National Science Foundation

Workshop Attendees and Content Contributors

Ahalt, Stan
Renaissance Computing Institute

Atkinson, Malcolm
University of Edinburgh

Baru, Chaitanya
San Diego Supercomputing Center

Bauzer Medeiros, Claudia
University of Campinas, Brazil

Borgman, Christine
University of California, Los Angeles

Bourne, Phil
Univ. of California, San Diego

Chen, Jake
Indiana Univ.-Perdue Univ. Indianapolis

Choudhury, Sayeed
Johns Hopkins University

Clark, Tim
Mass. General Hospital

Delany John
University of Washington

DeRoure, David
Oxford University

Dirks, Lee
Microsoft Research

Djorgovski, George
California Institute of Technology

Dozier, Jeff
University of California, Santa Barbara

Cynthia Dwork
Microsoft Research

Ellisman, Mark
Univ. of California, San Diego

Fox, Geoffrey
Indiana University

Goble, Carole
Manchester University

Goodman, Alyssa
Harvard University

Green, Daron
Microsoft Research

Heckerman, David
Microsoft Research

Hedstrom, Margaret
University of Michigan

Hunter, Jane
University of Queensland

Iwata, Shuichi
University of Tokyo

Joseph, Heather
Scholarly Publishing and Academic Resources Coalition

Kolker, Eugene
University of Washington

Konerding, David
Google

Langley, Pat
Stanford University

Lawrence, Bryan
British Atmospheric Data Centre

Lawrence, Katherine
University of Michigan

Li, Chung-Sheng
IBM

Liew, Chee Sun
National eScience Center, UK

Lin, Cui
Valdosta State University

Lynch, Clifford
Coalition for Networked Information

Michener, Bill
DataONE, Univ. of New Mexico

Mount, Richard
Stanford Linear Accelerator Center, Stanford Univ.

Nowell, Lucy
U.S. Dept. of Energy, Office of Science

Pike, Bill
Pacific Northwest National Laboratory

Pinkelman, Jim
Microsoft Research

Qin, Jian
Syracuse University

Salz, Joel
Harvard University

Szalay, Alex
Johns Hopkins University

Trefethen, Anne
University of Oxford

Van Ingen, Catharine
Microsoft Research

Viegas, Evelyne
Microsoft Research

Wang, May
Georgia Tech University

Xie, Dong
University of Oxford

Yarime, Masaru
University of Tokyo

Preface

The widespread use and deployment of cyberinfrastructure is transforming virtually every science and engineering discipline. Data volumes, computing power, software, and network capacities are all on exponential growth paths, and research collaborations are expanding dramatically. Scientific and educational endeavors are generating enormous stores of data from surveys, mobile and embedded systems, sensors, observing systems, scientific instruments, publications, experiments, simulations, evaluations and analyses. Scientists and citizens alike now routinely communicate by sharing data, software, papers, and visualizations. All sorts of publications are delivered in digital form; digital transmission and exchange drive social networks such as Facebook and Twitter, and they are fundamental for the rapidly growing importance of visual communication for both entertainment as well as for interpersonal relationships. Data are the basis of communication not only in science and education, but also in modern society.

With the realization of the central role that data play in science and society, we must address significant issues regarding support of science and education. These include the entire data-life-cycle from acquisition, curation, and storage of data, to access, manipulation and sharing. The more effectively that data can be manipulated, mined, managed, analyzed and served to communities, the better the conduct of science can be supported. The more we can eliminate boundaries in this exponentially growing sea of data, the better data can be shared enabling multidisciplinary and collaborative research, connecting the National Science Foundation (NSF) and its communities, and better coupling research and education. The more effectively students and faculty gain the data-intensive knowledge and skills, the larger the impact will be on science and society.

This is the world that the NSF ACCI Data/Visualization Task Force endeavors to address.

Dan Atkins, Tony Hey, Margaret Hedstrom

Table of Contents

Executive Summary

The Task Force strongly encourages the NSF to create a sustainable data infrastructure fit to support world-class research and innovation. It believes that such infrastructure is essential to sustain the United States' long-term leadership in scientific research and to capitalize on a legacy that can drive future discoveries, innovation, and national prosperity. To help realize this potential the Task Force identified challenges and opportunities that will require focused and sustained investment with clear intent and purpose; these are clustered into six main areas:

(1) **Infrastructure Delivery** - Acknowledge that data infrastructure and services are essential research assets fundamental to today's science and worthy of long-term investments. Make specific budget allocations for the establishment and maintenance of research data sets and services and associated software and visualization tools.

(2) **Culture and Sociological Change** - Introduce new funding models that reinforce expectations and institute specific conditions for data sharing. Create new norms and practices for citation and attribution so that data producers, software and tool developers, and data curators are credited with their contributions to scientific research.

(3) **Roles and Responsibilities** - Recognize that responsibility for data stewardship is shared among Principal Investigators, research centers, university research libraries, discipline-based libraries and archives, national scientific agencies, and commercial service providers. Determine a model for data stewardship and trust relationships among these parties in which there is clarity regarding ownership of data, software, and services, and a delineation of roles and responsibilities where interdependencies exist.

(4) **Economic Value and Sustainability** - Develop and publish realistic cost models to underpin institutional/national business plans for research repositories/data services.

(5) **Data Management Guidelines** - Identify and share best practices for critical areas of data management.

(6) **Ethics, Privacy and Intellectual Property** - Invest in the research and training of the research community in *privacy-preserving data-access* so that PIs can embrace privacy by design.

The Task Force believes that focusing on and investing in these challenges and opportunities will drive transformational research founded on cyberinfrastructure-enabled science. These recommendations stand to improve repeatability and reproducibility of science. They will increase access to data needed by individual investigators and small projects that constitute the "long-tail" of the scientific research community. Over time, investments in infrastructure for data analysis, reuse, and archiving will create jobs in software and tool development, data storage, and data curation, and increase the development pipeline of individuals with these key skills. Such skills will help the U.S. develop its educational pipeline and develop world-class expertise in large-scale data management, curation, and analysis - all critical for increasing scientific productivity, enabling scientific discovery and technical innovation.

The Task Force notes that its specific recommendations for improved data management should be viewed as a means to accomplish scientific discovery not as ends in themselves. Also, the Task Force acknowledges that a policy of retaining all scientific data is impractical and therefore recommends that the NSF support researchers and research communities undertaking the effective triaging of data for retention, archiving, and deletion. Furthermore, there are important lessons to be taken from other studies on infrastructure projects and, in particular, the dynamics and interplay between people, technology, institutions, and data [(1)].

The Task Force has not identified any recommendations for specific investments in visualization. This topic was explored, and the need for visualization and other analytical tools was acknowledged as integral to data-based advances in science. We suggest that the NSF refers to the Task Force focused on *Cyber Science and Engineering* for policy guidance on future visualization-related investments/programs.

Introduction and Charge

Data availability and the seamless access to scalable cyberinfrastructure are set to characterize scientific innovation and discovery in the upcoming decades and arguably even well beyond. Moving to a data-rich scientific context requires a shift in focus for NSF as it creates and sustains data services, as it builds scalable and open cyberinfrastructure, as it supports the long-tail of scientific endeavor, and as it improves access to and involvement in science by today's "citizen scientists."

The NSF report, *Cyberinfrastructure Vision for 21st Century Discovery*, describes the changing nature of science and engineering with the following words [(2)]:

"Science and engineering research and education have become increasingly data-intensive as a result of the proliferation of digital technologies, instrumentation, and pervasive networks through which data are collected, generated, shared, and analyzed. Worldwide, scientists and engineers are producing, accessing, analyzing, integrating, and storing terabytes of digital data daily through experimentation, observation, and simulation. Moreover, the dynamic integration of data generated through observation and simulation is enabling the development of new scientific methods that adapt intelligently to evolving conditions to reveal new understanding. The enormous growth in the availability and utility of scientific data is increasing scholarly research productivity, accelerating the transformation of research outcomes into products and services, and enhancing the effectiveness of learning across the spectrum of human endeavor."

The NSF's vision for cyberinfrastructure calls for the development of a national-level system of hardware, software, data resources, and services to "enable transformative discoveries in science, support national competitiveness, and to take full advantage of the wide availability of unprecedented volumes of rich data." The Office of Cyberinfrastructure at the NSF has established an Advisory Committee on Cyberinfrastructure (ACCI) to advise all of NSF on cyberinfrastructure implications and requirements. The Committee established a number of Task Forces to consult with the community on key aspects of cyberinfrastructure and this report is concerned with aspects of Data and Visualization.

The charge to the Task Force on Data and Visualization was to:

- Assess development of systems, methods, and policies related to data collection, analyses, management, and storage applications that are enabling for multiple disciplines.
- Establish an advocacy basis for a new foundation of cyberinfrastructure capabilities to enrich discovery, integration and accessibility of data, computational tools, and preservation capacity for the future.

The initial chair of the Task Force was Professor Shenda Baker, but the concluding phase of the work of the task force was led by co-chairs Professors Dan Atkins and Tony Hey, together with Professor Margaret Hedstrom.

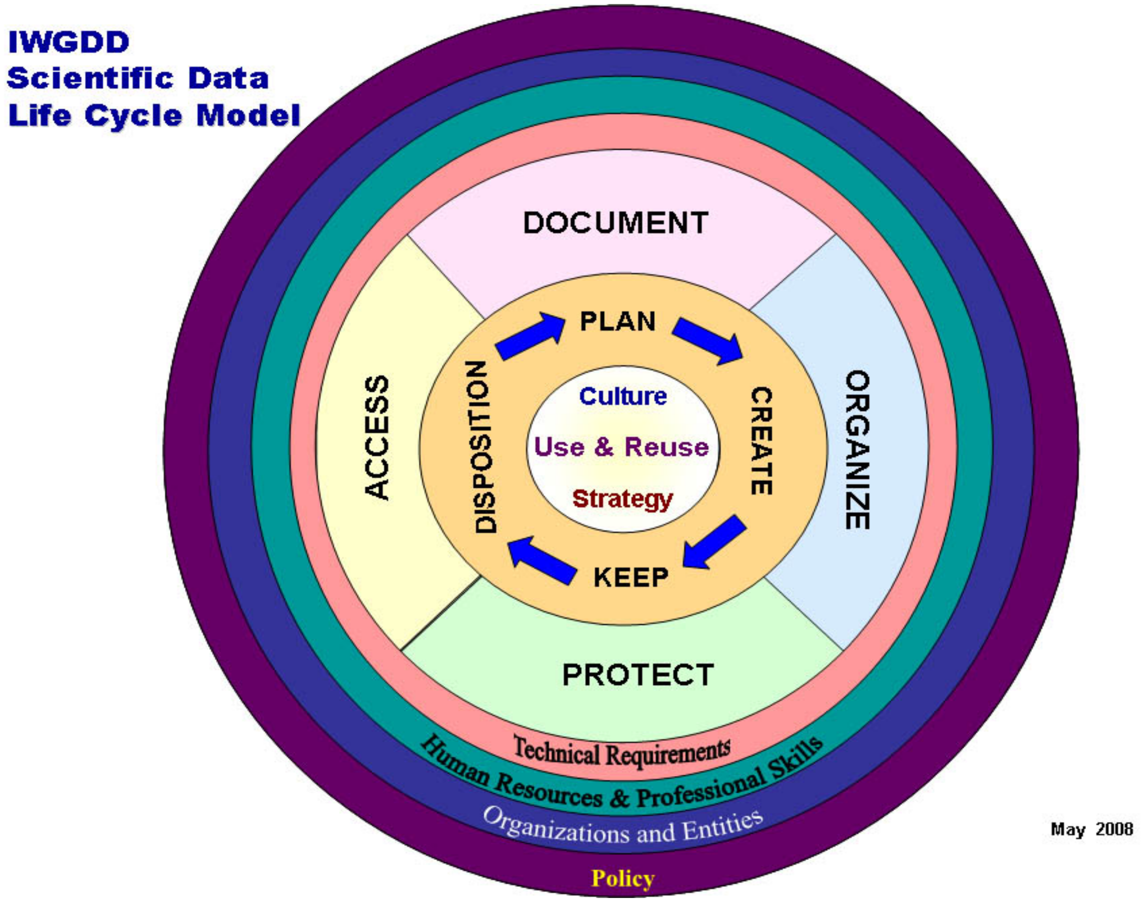
Previous reports have covered closely related topics dating back over a decade and some of these are listed in the Selective Bibliography. Of particular note are the 2005 report of the National Science Board on *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* [(3)]; the 2008 report of the NSF Task Force on Cyberlearning, *Fostering Learning in the Networked World: The Cyberlearning Opportunity and Challenge* [(4)]; the 2009 report of the Interagency Working Group on Digital Data, *Harnessing the Power of Digital Data for Science and Society* [(5)]; and the 2010 report of the Blue Ribbon Task Force on *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information* [(6)]. Additional relevant material can be found in the recent workshop report for the Mathematical and Physical Sciences (MPS) community *Data-Enabled Science in the Mathematical and Physical Sciences* [(7)] and a recent white paper on *Data Intensive Science in the Department of Energy* [(8)] giving a DOE perspective on the opportunities and needs of data intensive science.

The Task Force recommendations in this report have been informed with the benefit of such previous analysis and recommendations, and these and other related reports have provided useful context and some helpful reference points for this report. In particular, the NSF-ACCI Task Force Report on Grand Challenges includes a section on "Data and Visualization" and this Task Force endorses their brief recommendations on Data Visualization in section 6.4 of their report. This is why, although the subject was considered and discussed by this Task Force, there is little explicit emphasis on visualization in this report to avoid duplication of effort. There was consensus in this Task Force that data and visualization are interdependent and both are essential ingredients of data-driven science.

Many of the prior discussions of data and visualization issues in e-science have taken place within the main constituencies for advanced cyberinfrastructure: 1) computer scientists and engineers who are building the next generation of CI capabilities, 2) scientists who want to leverage new CI capabilities with the deluge of data to advance data-intensive science, and 3) librarians, archivists, data curators, and other information professionals who are developing new services to satisfy the demand for access to scientific data scholarly communications. Our work is distinguished from previous efforts in that the workshops targeted leaders in all three of these communities and the recommendations reflect a broad consensus of the participants from multiple perspectives.

Data Intensive Science

The vision of the Interagency Working Group on Digital Data is one of a "digital scientific data universe in which data creation, collection, documentation, analysis, preservation, and dissemination could be appropriately, reliably, and readily managed, thereby enhancing the return on our nation's research and development investment by ensuring that digital data realize their full potential as catalysts for progress in our global information society." Their report uses a model of the Digital Data Life Cycle to introduce and categorize the roles fulfilled by several types of organizations and communities. Although data analysis



and data processing has traditionally been associated with computational centers for so-called “big science” and individual investigators for “small science”, increasingly other types of organization such as libraries and archives are playing an important role in the data life cycle. Support entities exist across many sectors, including government laboratories, universities, and other research institutions, both commercial and non-profit. In addition, new specializations in data tools, infrastructure, and management are emerging as a result of the need to support all aspects of the data life cycle. These include the data scientists, themselves, together with digital curators and archivists who are expert in the techniques and technologies required for data preservation and re-use.

Data-intensive science consists of three basic activities: capture, curation, and analysis. Funding is needed to create a generic set of tools that covers the full range of activities—from capture and data validation through curation, analysis, and ultimately permanent archiving. Data curation covers a wide range of activities, starting with finding the right data structures to map into various stores. Curation also includes the schema and the necessary metadata for longevity and for integration across instruments, experiments, and laboratories. Without such explicit schema and metadata, the interpretation is only implicit and depends strongly on the particular programs used to analyze it. Ultimately, such uncurated data are as good as lost, even if the bits are stored forever, because they cannot be interpreted correctly. We must therefore think carefully about which data should be able to live forever and what additional metadata must be captured to make this feasible.

Data analysis covers a whole range of activities throughout the workflow pipeline, including the use of databases (versus a collection of flat files that a database can access), analysis and modeling, and data visualization. Although most of the major data-intensive research projects typically have a significant element for software, the vast majority of researchers in smaller projects have to make do with the widely available, but increasingly inadequate software tools. In the future, these scientists will need more powerful and sophisticated software tools to mine, analyze, visualize, and organize their data sets.

Scientific communication, including peer review, is also undergoing fundamental changes. In some fields, most notably physics and the life sciences, researchers are finding traditional forms of scholarly publication too slow. Research libraries face difficulty keeping up with rising journal prices. With the increasing size of scientific data sets it is becoming more and more difficult to find and access the data needed to reproduce the scientific analysis in a scholarly paper. Ultimately, such a trend could undermine the whole validity of the scientific record and hamper science’s ability to progress without wasteful duplication of effort. Fortunately, new forms of scholarly communication are emerging to address some of these problems. Research sponsors, journal editors, and scientists are requiring deposit of data with publications so that results can be reproduced. Many fields are starting public open access journals as an alternative to expensive commercial databases. Some research libraries are extending their traditional role as stewards of published scholarly output to embrace scientific data as well. These initiatives need to be monitored for emerging best practices, nurtured and spread to more institutions and disciplines, and coordinated so that the current *ad hoc* collection of largely disconnected small-scale and short-term initiatives become reliable components of data infrastructure.

For further background, a more complete exploration of the issues associated with data-intensive science is contained in *The Fourth Paradigm: Data-Intensive Scientific Discovery* (9) and summarized by the *Harvard Business Review* article "The Next Scientific Revolution." It should be noted that such a dramatic change in sciences has already been recognized as opportunity for competitive advantage as discussed in the European Commission's report *Riding the Wave: How Europe can gain from the rising tide of scientific data* (10).

Citizen Science

If people do not understand what a cell is how can they understand the ethics and implications of stem-cell research? Without an understanding of molecules and DNA how can the public understand the principles of heredity and risks in healthcare and disease management? Or, put another way, scientific illiteracy undermines citizens' ability to take part in the democratic process (29). NSF can catalyze community engagement in exciting scientific discovery and, through this, both advance scientific discovery and help educate U.S. citizens in key scientific principles.

There are now many examples of meaningful citizen science engagement. Galaxy Zoo activities, for example, give a useful indication of the latent appetite for scientific engagement in society. This is a collection of online astronomy projects that invite members of the public to assist in classifying galaxies. In the first year, the initial project boasted over 50 million classifications made by 150,000 individuals in the general public and quickly became the world's largest database of galaxy shapes. So successful was the original project that it spawned Galaxy Zoo 2 in February 2009 to classify another 250,000 SDSS galaxies. The project included unique scientific discoveries such as Hanny's Voorwerp (28) and 'Green Pea' galaxies.

World-class data infrastructure for world-class research

The Task Force strongly encourages the NSF to create a sustainable data infrastructure fit to support world-class research and innovation. It believes that such infrastructure is essential to sustain the U.S.'s long-term leadership in scientific research and drive future discoveries, innovation, and national prosperity. To help realize this potential the Task Force identified challenges and opportunities that will require focused and sustained investment with clear intent and purpose. Unlike previous reports we are specifically including statements of the underlying issues/concerns as these provide essential context for the Task Force recommendations. The recommendations naturally cluster into six main areas:

(1) Infrastructure Delivery;

(2) Cultural and Sociological Change;

(3) Roles and Responsibilities;

(4) Economic Value and Sustainability;

(5) Data Management Guidelines;

(6) Ethics, Privacy and Intellectual Property.

In addition, where relevant, examples of leading practices are provided to show the type of research or outcome that can be realized with appropriate policies, investment strategies, and measurement systems. Although these six areas are in many ways interrelated, we believe that it is helpful to factor the discussion in this way to make the causalities in the underlying issues clearer and the potential opportunities easier to identify. Those clarifications, in turn, allow easier characterization of the key enablers for the improved outcomes. In each area, the Task Force suggested one or two key recommendations which, with limited resources and constrained budgets, should be considered as the highest priority and those most likely to affect the creation of a world-class cyberinfrastructure capable of supporting world-class science.

(1) Infrastructure delivery

Moving from technical experimentation to reliable infrastructure: The Task Force believes the requirements for the sustainable development, delivery, and maintenance of long-term data infrastructure are currently, frequently confused with those of technical experimentation. This would be challenging in itself, however the situation has been compounded by costly duplication of efforts with no substantive coordination or collaboration between groups developing essentially the same type of experimental software infrastructure. Such a situation is clearly not sustainable and the Task Force believes that U.S. scientific research is suffering as a result. It should be noted that this problem has been recognized in other countries and by other funding agencies – most notably in the U.K. which was early to spot the potential issue and, to mitigate the risk, created the U.K.'s Open Middleware Infrastructure Institute (OMII) with the focus of providing and supporting critical e-science software infrastructure relevant to its local e-research community. By doing this the U.K. was able to prevent much of its e-science investments from being used to spawn multiple variants of the same basic software infrastructure. Instead, this approach enabled the funding to be used to solve a software problem once for the benefit of the whole e-science community.

The Task Force has no hesitation in encouraging the NSF to overcome this challenge and make a clear commitment to the sustained delivery and maintenance of robust, data-centric cyberinfrastructure services. The Task Force recognizes the deep complexities and challenges in this statement but firmly believes that a committed and sustained investment in data services is overdue. This effort will require some funding to be committed on a longer-term basis than traditional three-year NSF grants. The Task Force also recognizes that there is important overlap with the other Task Force recommendations (e.g. Software and Computing Task Forces) and that coordination of recommendations and follow-up will be required.

Preserving data to preserve the planet

The era of human-made objects orbiting Earth began with the launch of Sputnik in October 1957, and much has been written about Sputnik's beneficial stimulus to U.S. science. The U.S. satellite Explorer 1 followed in January 1958 and made the first scientific discovery from space about our planet, namely that Earth was surrounded by powerful radiation belts later named the Van Allen belts. Since then, satellite observations have been a fruitful component of data-intensive science, continuing to lead to unexpected discoveries about the universe and about processes that shape Earth's environment.

The story of ozone depletion illustrates how laboratory experiments, surface observations, and the capture and retention of data from many satellites led to a major milestone in mankind's understanding of its impact on the environment – in this case the Montreal Protocol, an international agreement to phase out ozone-destroying, anthropogenic, halogen-containing compounds.

In 1974 Sherwood (Sherry) Rowland and his graduate student Mario Molina published a laboratory study demonstrating that chlorofluorocarbons, a class of halocarbons, catalytically reduce stratospheric ozone formation in an environment of ultraviolet light. The science community urged the elimination of such ozone-destroying compounds because ozone protects Earth's inhabitants from damaging ultraviolet radiation. While halocarbons were soon largely eliminated as propellants in aerosol sprays, industry fought their banning as refrigerants and other applications.

In 1984, field and satellite observations revealed the magnitude and extent of ozone depletion. The Antarctic ozone hole was first identified in upward-looking UV radiation measurements by the British Antarctic Survey, and examination of data from the TOMS instrument (Total Ozone Mapping Spectrometer) confirmed the continental scale of the ozone hole, its annual appearance in the austral spring starting in the early 1980s, and its progressively increasing size. This unexpected magnitude in the depletion of ozone and the size of the ozone hole suggested that the stratosphere still held surprises. Subsequent measurements from a variety of satellites identified concentrations of trace species that lead to or catalyze ozone destruction. Chlorine from the breakdown of chlorofluorocarbons forms relatively inert hydrochloric acid (HCl), which reacts on the surface of ice crystals in polar stratospheric clouds (also discovered from satellite observations) to produce chlorine monoxide (ClO) that catalytically destroys ozone. Satellite observations also confirmed the presence of bromine monoxide (BrO), which is involved in reactions that are even more destructive of stratospheric ozone.

Satellite observations continue to track the size and depth of the Antarctic ozone hole and the more subtle, but dangerous, losses of ozone over heavily populated regions. Recent satellite observations show a decrease in halocarbons and the apparent beginning of an ozone recovery, increasing confidence that the Montreal Protocol is indeed achieving its goal.

Sherwood Rowland and Mario Molina (U.S.) and Paul Crutzen (Germany) shared the 1995 Nobel Prize in Chemistry in recognition of their work on the mechanics of ozone destruction.

This work is just one aspect of climate change which is now being added to through diverse data sources such as ice-core analysis, dendrochronology, and even careful interpretation of ships logs to help build an improved picture of the changing cycles of planet. This latter case is particularly remarkable since the retention of wind speed, temperature, and pressure data for vessels in the British Navy for the past 200 years is now providing a unique source of climate data and represents an incredibly far-sighted policy of data capture and curation[30]

The primary motivation for this investment should be to support U.S.-based research breakthroughs and innovation. However, it is important that this be done within a global context and with sensitivity to the broader research community. More specifically, a number of the existing research projects are undertaken across international boundaries and involve collaboration with leading scientists throughout the world. Some thought will therefore need to be given to ownership and access to data from remote collaborators and, perhaps, allowing parts of the data storage to be federated internationally to respect local data policies and regulations.

Robust data service infrastructure for breakthrough science: The existence of a robust and scalable set of scientific data services

- significantly accelerates startup of new projects (and the dynamic expansion of existing programs) by removing the start-up and lead-time for capital investments and the creation of one-off data services. This will allow for a more responsive and creative research context;
- promotes broader competition for service providers to provision services for users;
- avoids valuable graduate student resources from being diverted to act as departmental system administrators and/or software developers;
- provides opportunity for commercial innovation. As an example, consider the numerical software company NAG (Numerical Algorithms Group) which started in the U.K. but now has a U.S. subsidiary. NAG was started several decades ago by the numerical algorithm research community to curate and distribute mathematical/numerical software. One could envisage equivalent potential for data-service related commercial spin-offs.

Task Force Recommendations:

- **Key Recommendation:** Recognize data infrastructure and services as essential research assets fundamental to today's science and as long-term investments in national prosperity. Make specific budget provisions for the establishment and maintenance of data sets and services and the associated software and visualization tools infrastructure.
- **Supporting Recommendation:** Serve scientific communities' data service requirements through:
  - having key research domains identify and triage their essential data (including meta data) that needs to be retained and archived. Key questions include: "What data should we be gathering/curating today and what subset of that data will we need in 20-30 years' time? Which scientific researchers in other fields will need the data?"
  - issuing an open call for large-scale data services across these science disciplines and across a range of data types. This should be formulated to attract competition between third-party service providers (including partnerships with the private sector). If undertaken at a sufficient scale, new and highly favorable price-points for data services/storage could be achieved. Such a service should NOT exclusively focus on large-scale or what could be referred to as "petabyte data" but rather include mid/small-sized research where investigators have terabyte-sized data sets. It should also include ongoing data access and curation service levels that require and incent successful service providers to migrate data over time to new and more cost-effective media.
  - working with the research community to positively and actively promote open access to these new data services.

Examples:
The following provide examples of existing data archives and/or services promoting open access to scientific data services and tools.

- Founded in 1984 with support from the NSF, IRIS [(11)] (Incorporated Research Institutions for Seismology) is a consortium of over 100 U.S. universities dedicated to the operation of science facilities for the acquisition, management, and distribution of seismological data. IRIS programs contribute to scholarly research, education, earthquake hazard mitigation, and verification of the Comprehensive Nuclear-Test-Ban Treaty. This is an essential research data archive offering tools and tutorials for data access to researchers and promoting a broader understanding of seismology to the science community and the general public. IRIS's mission is to:

— Facilitate and conduct geophysical investigations of seismic sources and Earth properties using seismic and other geophysical methods.
— Promote exchange of geophysical data and knowledge through use of standards for network operations, data formats, and exchange protocols, and through the pursuit of policies of free and unrestricted data access.
— Foster cooperation among IRIS members, affiliates, and other organizations in order to advance geophysical research and convey benefits from geophysical progress to all of humanity.

- The National Institutes of Health support both the GenBank and Protein Data Bank databases, which have transformed the availability of and scientists' access to nucleotide sequences and protein structures. GenBank started at Los Alamos National Laboratory in 1982 as a database for nucleotide sequences, funded by several federal agencies. In 1992, GenBank transitioned to the newly created National Center for Biotechnology Information (NCBI). The Protein Data Bank for macromolecular structural data originated at Brookhaven National Laboratory in 1971 and was transferred to the Research Collaboratory for Structural Bioinformatics (RCSB) in 1999. The RCSB PDB is funded by several agencies including the NSF and the NIH National Library of Medicine (NLM) and other NIH Institutes. GenBank and the Protein Data Bank clearly demonstrate the huge value in providing open access to well-archived and structured scientific data and are core services for all current molecular biology.

— GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. There are approximately 107 million bases in 108 million sequence records in the traditional GenBank divisions and 150 million bases in 50 million sequence records in the WGS division as of August 2009. The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced at National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration, or INSDC. GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank continues to grow at an exponential rate, doubling every 18 months. GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.
— Established in 1971 at Brookhaven National Laboratory, the Protein Data Bank originally contained 7 structures. This archive is now the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. These are the molecules of life that are found in all organisms including bacteria, yeast, plants, flies, other animals, and humans. Understanding the shape of a molecule helps to understand how it works. This knowledge can be used to help deduce a structure's role in human health and disease, and in drug development. The structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome. The PDB archive is available at no cost to users.

- Inter-University Consortium for Political and Social Research (ICPSR) (12) is an international consortium of approximately 700 academic institutions and research organizations. With diverse funding sources such as NIH, NSF, Institute of Museum and Library Services, National Institute on Drug Abuse, Library of Congress, Department of Justice, and many other organizations, ICPSR provides leadership and training in data access, curation, and methods of analysis for the social science research community.

— It maintains a data archive of more than 500,000 files of research in the social sciences. It hosts 16 specialized collections of data in education, aging, criminal justice, substance abuse, terrorism as well as other fields. CPSR runs educational activities and provides curricula in research design, statistics, data analysis, and social methodology. It also leads several initiatives that encourage use of data in teaching, particularly for undergraduate instruction.
— The consortium has sponsored research that focuses on the emerging challenges of digital curation and data science.
— ICPSR maintains a number of significant thematic collections. These include:

- **Child Care and Early Education Research Connections** (RC): a comprehensive, easily searchable collection of more than 15,000 resources from the many disciplines related to child care and early education. The site offers the most current publications, as well as links to child care policy statements.
- **Health and Medical Care Archive** (HMCA): preserves and disseminates data collected by research projects funded by The Robert Wood Johnson Foundation, the nation's largest philanthropy devoted exclusively to improving the health and health care of all Americans.
- **National Archive of Criminal Justice Data** (NACJD): facilitates research in criminal justice and criminology through the preservation, enhancement, and sharing of computerized data resources. NACJD also promotes original research based on archived data and provides specialized training workshops in quantitative analysis of crime and justice data.
- **Resource Center for Minority Data** (RCMD): provides educators, researchers, and students with data resources so that they can produce analysis of issues affecting racial and ethnic minority populations in the United States. RCMD provides access and analytic tools enhancements to use of the vast array of available data. The archive collection allows researchers to track changes in outcomes and status of minority populations and contributing factors. Access to RCMD data is available to anyone at an ICPSR member university or institution.

Oceans of data

After a boating or aircraft accident at sea, the U.S. Coast Guard historically has relied on sea current charts and wind gauges to figure out where to hunt for survivors. But thanks to data originally collected by Rutgers University oceanographers to answer scientific questions about earth-ocean-atmosphere interactions, the USCG has a new resource that promises to literally save lives. **It is a powerful example that large data sets can drive myriad new and unexpected opportunities and it is an argument for funding and building robust systems to manage and store the data.**

There is a revolution underway in oceanography today. Scientists around the world are augmenting the ship-based expeditionary science of the last two centuries with a distributed, observatory-based approach involving instruments, facilities, and networked interactions with other scientists. These efforts, including those sponsored by the National Science Foundation (NSF) Ocean Sciences Division, are focused on routine, long-term measurement of episodic oceanic processes on a wide range of spatial and temporal scales. Such data are crucial to resolving scientific questions related to Earth's climate, geodynamics, and marine ecosystems. However, the same sensors and systems are also yielding valuable data that are being shared and repurposed for government and commercial uses, including energy planning, defense, and even real-time life-and-death challenges such as ocean rescue.

At Rutgers University's Coastal Ocean Observation Lab, scientists have been collecting high frequency radar data that can remotely measure ocean surface waves and currents. The data are generated from antennae located along the eastern seaboard from Massachusetts to Chesapeake Bay. The network was built bit-by-bit to answer very specific scientific questions, such as determining the precise physical river flows of the Hudson River when it empties into the Atlantic Ocean in order to track its impact on the marine food chain. However, over time the data, and this research group's willingness to share it, are also providing previously unobtainable information for an array of users.

The New Jersey Board of Public Utilities (BPU), for example, is interested in developing an offshore wind farm industry. It turns out that surface currents serve as a proxy for sea breezes that can be localized and measured with high accuracy. BPU is using this historical data to plan the placement of equipment and project the energy likely to be captured.

The Department of Homeland Security (DHS) realized that Rutgers' radar data also includes the echoes of ships. Not only can the historical data allow DHS to analyze past shipping patterns and detect changes, but the technology holds the promise for what is called "over the horizon" ship detection that cannot be collected any other way. For example, DHS is looking to use the data to focus security checks on vessels that have not reported their location.

Perhaps the most dramatic sharing involves the U.S. Coast Guard. The Rutgers' experiments sought to identify highly accurate, real-time ocean circulation patterns. "Now, instead of developing a search box that could be as big as the state, they can integrate our data and get actual currents and decrease search areas for survivors," explains Oscar Schofield, professor of Bio-Optical Oceanography at Rutgers. "That raises the probability of faster rescue and higher survival rates."

One of the group's frustrations today, unfortunately, is the lack of funding to design and support long-term preservation of data**. A large fraction of the data the Rutgers team collects has to be thrown out because there is no room to store it and no support within existing research projects to better curate and manage the data. "I can get funding to put equipment into the ocean, but not to analyze that data on the back end," says Schofield.**

(2) Cultural and Sociological Change

Entrenched culture is a roadblock to change in the practice of scientific research: Influential scientists/leading researchers in positions of authority are typically unwilling to change a system that has made them successful. Indeed, they will often actively resist change to maintain the *status quo*. The Task Force identified that existing research culture doesn't encourage or reward the right practices or behavior with regard to data management and sharing.

Few researchers place importance on or value the people involved in data management and/or data curation. This leads to there being inadequate career opportunities for those essential to the future of scientific research and no clear pipeline of expertise to support the required skills and resources.

NSF's opportunity to catalyze data sharing: Introduce incentives and mechanisms to improve data sharing practices where minor changes catalyze a big impact on behavior. Recognize that the new generation of scholars is comfortable and arguably even visionary with regard to sharing and collaboration. The NSF, in partnership with the research community, has the opportunity to stimulate and reward healthy risk-taking in terms of data sharing, access, and collaboration. Moreover, opening up data access provides new opportunity for research by the diverse group of small scale scientists and by citizen scientists, and it creates opportunities for innovation by business.

Task Force Recommendations:

- **Key Recommendation:** Introduce new funding models that have specific data-sharing expectations and support researchers in meeting data management and data sharing requirements imposed by research sponsors. For example:
  - — Institutional funds to be made available for data services that encourage or enable career advancement in academia based on good data management, data sharing, and open access practices;
  - — Agencies to expect/require data sharing as a condition of funding (not simply requiring a data management plan);
  - — Costs of data production, management, and data sharing to be included in proposals;
- **Key Recommendation:** Create new citation models in which data and software tool providers are credited with their data contributions and establish metrics that recognize open access policies and sharing.
  - — Encourage change in citation patterns to include a role for citations (e.g. to value activities such as "data provider/curator" and/or "software tool provider" alongside "data analyzer" or "computational modeler"), which can help create a credit market for data and software sharing.
  - — Encourage publication of data in a citable form before paper publication as advocated in initiatives such as SageCite [(13)].
  - — Create specific project metrics that assess and monitor effective availability and accessibility of data.
- **Supporting Recommendation:** Encourage a freedom of research information principle to be used, where possible, to ensure the availability and accessibility of key scientific data sets by researchers, society, and industry (perhaps under appropriate license terms). Clearly there are privacy and confidentiality issues but the Task Force believes it is important that new principles and social contracts be created to encourage data sharing and data access. Social networking tools could play an important role in building communities of practice and supporting collaboration.

Examples:

- The open data sharing through *Galaxy Zoo* [(14)]*,* Microsoft Research's *World-Wide Telescope* [(15)], Google's influenza study [(16)] and IBM's *Many Eyes* [(17)] provide excellent examples of how open access to scientific data delivers multiple potential benefits. Specifically these examples:

— enable a new type of scientific innovation through large-scale data access coupled with data analytics and visualization;
— increase data-access by the science's long tail of researchers who otherwise have access to only modest local computing infrastructure and typically have limited data management skills;
— improve the public understanding of and engagement in science leading to a better chance that society makes educated and informed decisions on government investments and research;
— excite young people about careers in science and engineering.

Rapid data sharing speeding: Quest for Alzheimer's biomarkers

The promise of speeding up vital biomedical research by better and faster sharing of data is becoming real in an innovative Alzheimer's disease research partnership between the private sector and the National Institutes of Health. Called the Alzheimer's Disease Neuroimaging Initiative (32) or ADNI, it was launched in 2004 specifically to improve clinical trials for the dread neurological condition. One reason Alzheimer's research is so difficult is that researchers lack good biomarkers to track disease progression. In fact, the disease still can only be definitively diagnosed by brain biopsy after death.

ADNI attacks that challenge in several ways. First, it combines data from several volunteer subject groups and several diagnostic methods, including spinal fluid analysis, magnetic imaging resonance (MRI) scans, and positron emission tomography or PET. These tests are periodically performed on 800 volunteers on the spectrum from completely healthy, through mild impairment, to patients with clinically diagnosed Alzheimer's. As some volunteers progress from healthy to mild impairment or impairment to full-blown Alzheimer's, the hope is to find biomarkers that can more faithfully track the progression of the disease.

Not only can the data from the 14 different centers involved in the initiative be combined and compared, but it is also highly significant that the data is typically made publicly available within a week of being collected. Those two factors are catalyzing the energy of neuroscientists both at those centers and around the world. Hundreds of scientists have made tens of thousands of downloads from the ADNI website, and, of several dozen papers that have so far been published using ADNI data, a significant number were authored by researchers who are not even directly funded by the project. Scientists say the rapid sharing is motivating them to analyze and publish data more quickly, and companies are incorporating the data into their clinical trials for promising treatments.

Increasingly, scientists believe that Alzheimer's disease pathology may be present 10 or even 20 years before the onset of dementia. Three studies published by non-ADNI researchers suggest they are zeroing in on biomarkers that may help identify which patients are likely to progress to Alzheimer's disease. That could be highly significant in a devastating disease where earlier detection is considered key to better treatments.

(3) Roles and responsibilities

Lack of clarity on data ownership/stewardship: Currently there is confusion and ambiguity over who has data ownership and stewardship responsibilities throughout the data lifecycle. For example, it is unclear who is accountable for important issues such as the reproducibility of science, data retention, and data accessibility. Current guidelines appear weak and suffer from little or no policing or enforcement and as a result there is little or no effective accountability.

Getting clarity and giving purpose: NSF, U.S. universities, research libraries, and principal investigators (PIs) should agree on shared responsibility for new transformative data-intensive science. The NSF can capitalize on the groundswell of interest and opinion in the community but such a change is unlikely to happen spontaneously without leadership from major funding agencies. If it chooses, the NSF can change the context and promote an ecosystem that values and invests in good Research Data Management practices and thereby transform the way in which scientific research is undertaken. The Task Force believes that the NSF can act as a catalyst to enable institutional data management as well as being a valued funder of trusted and sustainable data storage services (in particular when combined with effective Infrastructure Delivery [18]).

Task Force Recommendations

- **Key Recommendation:** Orchestrate discussions with PIs and leaders in research universities, federal scientific agencies, funding bodies, and research libraries to determine a model for data stewardship that define sets of data and software services, delineate roles and responsibilities, and attend to interdependencies among respective services. This model must clarify important principles such as who supports and provides long-term storage of key research data, what access policies there should be, what archival processes are assumed. Note that the identification of roles/responsibilities will benefit significantly from the recommendations in (5) Data Management Guidelines. However, the Task Force believes NSF must recognize that guidelines in themselves are NOT sufficient, and these need to be accompanied by clarity over roles and responsibilities [19].

LHC's data service roles

eGee
Enabling Grids for E-sciencE
Open Science Grid

Image courtesy of CERN

The Large Hadron Collider [33] is a well-cited example of how science instruments can generate enormous volumes of data. While many marvel at the engineering behind the instrument itself, the global data infrastructure is an equally important characteristic underpinning the research. In this case a *DataGrid* distributes petabytes of data from the *Tier 0* site at CERN to a network of *Tier 1* processing and archival sites throughout the world. This federated design is an essential component of the cyberinfrastructure and key to the international collaboration. Indeed, it is a critical feature of the new way in which High-Energy Physics (HEP) research is conducted. The Tier 1 sites have sufficient storage capacity for a large fraction of the data and round-the-clock support for the computing grid. In addition, there are more than 150 Tier 2 data/compute service centers active in the LHC experiment.

Of particular relevance to this report is the fact that there are dependencies between data, software, and computer service providers across the HEP community. They have specifically identified roles and clear service level agreements that bind the research community and research centers together. As a result, this community no longer considers scientific innovation to be the preserve of a single researcher or research team but rather a global collaboration anchored in data sharing.

- **Supporting Recommendation:** The NSF should actively review project Data Management Plans and more directly and intentionally monitor the actual level of data *openness*, *accessibility,* and level of effective *sharing* across the projects it sponsors. It should give clear feedback during project and program reviews of what is expected by PIs and the key project stakeholders and either withhold payment or restrict future sponsorship where poor data management and restrictive access practices persist. Specifically, projects should be encouraged to seek increasingly open data policies where this is possible.

RoI of scientific data services

One of astrophysics' great quests is to comprehend the mysterious "dark energy" which acts to accelerate the expansion of the universe. Our current understanding of dark energy comes primarily from the study of supernovae, which help measure that expansion. The Nearby Supernova Factory (31) (SNfactory) is an international astrophysics experiment designed to discover and measure Type Ia supernovae in greater number and detail than has ever been done before. It has about 30 members; about half in the U.S. and the other half in France. On any given night, the project's primary telescope, which is in Hawaii, is used to collect up to 80 GB of data and is typically operated by a geographically separated group of two to six people.

Astrophysicists collaboratively operating large telescopes on a limited schedule face extraordinary real-time challenges ranging from weather conditions to language and cultural differences. **However, from the start, data curation and management were considered a priority in this project. The result is that Sunfall (34), (SuperNova Factory AssembLy Line), a collaborative scientific data management and visual analytics system created to support SNfactory, is an example of the significant return on investment – both in terms of financial resources and in terms of scientific productivity -- that cyberinfrastructure can provide.** Sunfall brought together an interdisciplinary team including physicists, computer scientists, and software engineers; the system incorporates sophisticated astrophysics image processing algorithms, machine learning capabilities, and astronomical data distribution and analysis with a usable, highly interactive visual interface designed to facilitate collaborative decision-making.

Because all the collaborators were able to provide input, the team targeted what had been recognized previously as key data chokepoints. They also created a real-time chat system that allowed the team to process changing conditions and make decisions efficiently. As a result, the overall system significantly reduced several labor-intensive steps. **The new solution reduced false supernovae identification by 40%; it improved scanning and vetting times by 70%; and it reduced labor for search and scanning from 6-8 people working four hours per day to one person working one hour per day.** Not only did the system pay for itself operationally within 1.5 years, but it enabled new science discovery. It led to ten publications in 2009 in both computer science and physics journals, and three best paper awards in computer science

Cross-disciplinary teams and distributed work structure and management are becoming essential to some of the grand challenges of the 21st century. Sunfall demonstrates that a well-designed cyberinfrastructure that addresses these new challenges will be both essential and impactful.

(4) Economic value and sustainability

Little understanding of long-term storage costs and unsatisfactory methods of valuing research data: Naturally occurring data (as opposed to simulated or derived data) is invaluable as it often cannot be retaken or reproduced and in some cases the true value of data is only realized decades later. That said, it is unclear what the actual costs/value should be associated with long-term data management/preservation and there is no easy or agreed upon method by which to determine the opportunity costs from its losing/deleting/neglecting data and software assets.

Valuing research data and software assets: The NSF has the opportunity to provide leadership in developing sustainable service models (in concert with other stakeholders) that are a prerequisite for contemporary science. In doing this, the intent is to flag that data management and software tools (e.g. for visualization) are a necessary investment for breakthrough science and not a hidden cost. Through clarifying the economics at play in data and software management, the NSF can formally acknowledge these as viable upfront components of research. It can also actively encourage increased sharing and data and software reuse and thereby improve the overall return on investment (RoI) for its programs. Clearly, ensuring data and software sustainability reduces potential cost of having to repeat research and/or duplicate development costs [20].

Task Force Recommendations

- **Key Recommendation:** To develop and publish realistic cost models to underpin institutional/national business plans for research repositories/data services, the NSF should:

  — Commission studies that evaluate data management, storage, access and retrieval costs models; it should also establish costs models for software development and maintenance. These models should support the research community by characterizing the expected cost profile for management and preservation of its research software and data. The models must be able to handle multiple data modalities and access patterns across different sizes of research communities. These models will allow researchers to plan for long-term success by better anticipating their data and software cost profile. Such an undertaking will require the NSF to update the models to monitor and evolve its policies to meet emerging needs/trends.
  — Work with the private sector and accept the role of market forces. NSF, perhaps in collaboration with other agencies, could negotiate “bulk” service agreements on behalf of its research community. Such co-ordination has the potential to drive volume discounts and an advantageous pricing for third-party services. Although initially suggested in relation to data storage/management such a model can apply to software development for which professional third-parties are commissioned to undertake the development and documentation of commonly-required software tools across the scientific research community.
  — Measure success by data and software reuse and by project proposals adopting these models in the data and software funding requests. Long-term success will be evident in the sustained availability of key research data and software assets.
  — Encourage each major scientific domain to triage its major data resources to identify which can be sustainably maintained and for which there is a solid case for long term RoI.
  — Require that these or equivalent sustainable models/services be referenced and used as a component of the Data Management Plan for any large-scale data collection or software development projects.

- **Supporting Recommendation:** The NSF should investigate data and software licensing options with a view to helping supplement research budgets. This can be done by encouraging the following:

  — New data licensing agreements
  — Specific articulation of rights/licenses as part of any funding/investment program
  — Reuse of non-PI data for commercial services (with license revenues)
  — Adoption of an appropriate license such as the Creative Commons Zero Waiver (CCZero) [21] or the Public Domain Dedication and License (PPDL) for use within scientific research community as suggested, for example, in the Panton Principles [22].

- **Supporting Recommendation:** There is potential for business value derived from both data and from the software developed as part of the NSF's research investments. The Task Force recognizes that the prior two recommendations could include provision for third-party revenue streams, which could defray some of the costs for data preservation. More specifically, there is the potential for commercial value to be derived from some of the research data. The research community can therefore supplement its NSF funding through careful stewardship and appropriate licensing of these resources. Currently there are no clear guidelines, recommendations, or implicit assumptions as to how data or software will be exploited by the commercial sector. Careful consideration of the potential could result in shared costs for data preservation and/or more money being made available for scientific research.

Examples:

Longitudinal studies have huge and measurable value and clearly represent critical resources for future research:

- Climate change data;
- National census data [23].

(5) Data management guidelines

What does 'good' look like? Data management best practices are not well understood among most scientific researchers. This is in part because leading practices have not been sufficiently well identified but also because existing effective approaches and successful solutions are not well promulgated through the scientific community. This is perhaps understandable given that there is only weak consensus over how best to capture, triage, annotate, store, provide access to scientific data. Moreover, many if not most scientists focus on the shortest path to a particular scientific result rather than the best long-term solution for data reuse or data-service provision to the broader research community.

Establishing examples of leading data management practices: Although the NSF does currently require prospective project proposals to provide statements regarding their data policies and the intended approach to data management, there is a significant opportunity for a more prescriptive set of guidelines/leading practices to be established. Put simply it's time to define "what good looks like." In this way NSF can foster better data management practices within its project portfolio and effectively encourage/reward improvement in data management practices.

Task Force Recommendations:

- **Key Recommendation:** Identify and share best-practices for the critical areas of data management including:
  - — the overall economics and resource requirements of good data management practices (see specific recommendation on economics and sustainability);
  - — methods/approaches to data triaging to help research communities decide what to retain and, perhaps more importantly, what to discard;
  - — capture and management of metadata;
  - — solutions for effective data archiving including how to handle "bit rot";
  - — tools and techniques for data migration;
  - — how to effectively and securely offer data services/access to various stakeholder communities such as: collaborators, other researchers in the scientific domain and, where appropriate, the general public;
  - — approaches for avoiding *orphaning* of data (cf. equivalent issue with orphaned software identified by the NSF-OCI Task Force on Software for Science and Engineering);
  - — how to associate scientific publications with the underlying data and software assets (to improve the reproducibility of science).

The NSF should share examples of these leading practices and, in particular, visibly celebrate and reward effective data management. For example, the NSF could focus its funding on supporting data centers that effectively work with PIs to achieve long-term retention and access.

- **Supporting Recommendation:** The Task Force recommends the NSF consider an initial focus on mid-scale science (e.g. neutron scattering, high-power NMR, high-volume light source data) as there is a large volume of science data that is currently being lost through inadequate focus on data management. Typically these fields have too much data for low-cost local solutions but are not of sufficient size to warrant large-scale data infrastructure.
- **Supporting Recommendation:** To aid in this the NSF could broker PI-data center relationships and recommendations (i.e. given certain requirements and conditions, PI A should consider data retention services X, Y, and Z.)
- Specifically encourage data *reuse* for research and have this as a focus of some proportion of the science investments.
- Require more complete data management policies for project proposals, in particular, requiring data retention and open sharing plans in grant submissions

Examples:
See inset on 'U.K.'s Digital Curation Centre - Leading in Digital Curation and Data Management Practices'.

(6) Ethics, Privacy and Intellectual Property

Expanding research access to data while enabling data privacy is still an unsolved problem: The Task Force sees that governments, universities, and commercial companies often struggle with privacy issues when attempting to increase or extend researcher access to data. Specifically, issues reside in the act of granting data access while, at the same time, protecting the confidentiality of the data and/or handling the business sensitivity of the data and/or dealing with the privacy of the individuals whose data are recorded. In all such cases protection implies the avoidance of unethical or inappropriate usage. The growth in cyberinfrastructure raises new and far more challenging questions about the protection of privacy associated with electronic databases involving individuals, families, animals, plants, and other groups, as well as of organizations. One can observe therefore that data has a series of sensitivity taxes attached to it that can make it difficult to share:

Loyalty Programs with Data Privacy and Data Ethics

An interesting general example of strong participation from members of society in a program that involves providing personal data can be found in hospitality loyalty programs such as Frequent Flyer programs.

A study done on hotel loyalty showed that high participation in the program was due to the benefits returned to users - while in fact users have a high level of concerns about privacy [27].

To address users' concerns, such programs use a *systems thinking* approach with ethical principles. This takes into account a broad context by including the customers, the users who must work with the data, business goals, as well as the interactions among these concerns.

Adopting a systems orientation and considering three ethical principles (such as minimizing harm, offering respect, and operating consistently) seems to reassure hospitality loyalty program customers that their data are secure.

Although taken from a business setting, this approach suggests similar methods may be applicable for handling sensitive data in a research setting.

- Privacy – there are privacy concerns, for example, in the sharing of medical data or consumer preferences collected for one specific purpose but then repurposed.

- Legal and Business – there are legal and business concerns about who has ownership of certain data, such as in potential infringement of copyright laws;

UK's Digital Curation Centre:
Leading in Digital Curation and Data Management Practices

Creation of the U.K.'s Digital Curation Centre [35] (DCC) was a key recommendation in the Joint Information Systems Committee (JISC) program in its Continuing Access and Digital Preservation Strategy, which argued for the establishment of **a national center for solving challenges in digital curation that could not be tackled by any single institution or discipline**. DCC's mandate includes the provision of generic services, some development activity, and research. During its Phases 1 and 2 (March 2004-February 2007 and March 2007-February 2010), the DCC had a target group defined as those engaging in digital preservation and curation activities within U.K. Higher and Further Education. This group included data specialists, records managers, librarians, archivists, researchers (as data creators), and policy-makers. The DCC also sought to engage in project activity with the public and commercial sectors, international sister organizations and standards working groups, recognizing that the advancement of tools and processes for digital curation depends on developments that take place beyond the U.K's higher and further education sector as well as within it. Hence, establishment of the DCC Associates Network became a forum for cross-sectoral communication on important problems more broadly.

By the start of Phase 2, the emphasis of DCC activity had shifted considerably towards increased and **direct involvement with the active research community**, as exemplified by the creation of an e-Science Liaison function and the conduct of immersive discipline case studies by the SCARP project**.**

Phase 3 (March 2010 - February 2013) has brought the introduction of further structural changes, with a shift away from the development of curation tools and a renewed focus on building capacity, capability and skills for data curation across the U.K.'s higher education research community. This new emphasis has exposed a critical dependency upon the contribution of a network of practitioners beyond the core DCC, who will be crucial to the exponential growth of effective data curation practice.

The DCC Phase 3 team, with its core at the University of Edinburgh and its partners at UKOLN (University of Bath) and HATII (University of Glasgow), now concentrates on the provision of **expertly mediated access to resources, originating both from the DCC and elsewhere; an advocacy and community development program designed to produce a nationally coherent movement for change**; all underpinned by a training program aimed at nurturing the transfer of knowledge and best practice between data producers, users, and custodians.

DCC has produced a set of guidelines for U.K. researchers creating management data plans. **It has issued templates and guidance on how to think about data curation and how to go about considering the policy decisions**. These guidelines are now used widely by researchers around the world.

value matters of national security or law enforcement over their own stated privacy policies, or where information itself may lead to negative societal consequences (e.g. ability to manufacture weapons , synthesize hazardous chemicals or snoop on private citizens);

Currently there are no domain independent technical solutions for sharing data in a secure way. Moreover, what it means to “share” data seems to depend upon context. For example, privacy may be compromised even if raw data are not shared at all, while copyright laws perhaps may not be broken when copyrighted material is summarized.

Any potential solutions need to take into account the tension between all the protection taxes above. These solutions need to be technically robust (e.g. access to information via privacy-preserving mechanisms instead of access to raw data) and provide incentives for individuals to participate in what should be a “safe” open data world.

Task Force recommendations
The Task Force identified two areas NSF can focus on in the near term:

- **Key Recommendation:** Increase investment in research and training of the research community in privacy-preserving data-access so that PIs can embrace privacy by design with clear guidelines on producing a privacy data plan. A concrete example may be for NSF to sponsor the building of a library of "stable" (high accuracy) differentially private implementation of standard data analysis tools to be used as part of the privacy data plan. Increasing research effort in formalizing/modeling ethics to describe the “societal data contract” a society needs to abide by in open data world. A concrete example may be for NSF to sponsor studies on data-driven ethics with a *systems thinking* orientation and identification of ethical principles in a couple of domains e.g. patient or finance data; see ‘Loyalty Programs with Data Privacy and Data Ethics’ inset.

- **Supporting Recommendation:** Explore and establish new data licensing mechanisms. For instance, the redefinition of IP ownership for major discoveries funded by public funds could be a means to incentivize PIs to make their data accessible in a way that will have the characteristics of a “privacy by design” data set and can fit the “societal data contract.” In particular, NSF could look into allowing restrictive use of IP, with earnings going back to the research pool to drive more innovations. Specifically, the ‘Guide to Open Data Licensing’ [(24)] from the Open Knowledge Foundation provides some useful context and examples of current thinking in this domain.

Examples
Today it is easier to find examples of risk associated with failures of privacy, ethics, and IP protections than exemplars implementing robust technical and societal solutions allowing scientists to successfully share data for research (be it raw data or access done via privacy-preserving mechanisms). A few examples of these risks include the following:

- AOL’s release of user search data leads to PII exposure [(25)].

- Patient record information allows 1997 governor of Massachusetts to be re-identified using only his date of birth, gender and ZIP code [(26)];

This state of affairs has significant consequences on both the development of science as a discipline (which requires reproducibility of experimentation) and on reducing data-driven breakthrough innovations in the academic world (which requires access to data).

For example, data mining of comprehensive electronic medical records has the potential to detect causes of rare diseases, but this data cannot be made available for research until the recommendations suggested in this report are addressed.

Historical satellite data delivers high 'VOI'

Good policy decisions demand good data, nowhere more so than in the complex regulatory and planning considerations that attend land-use decisions with environmental impact. Uncertainty can spark either over- or under-regulation, and either scenario can have negative societal and economic impacts. Although the expense of maintaining publicly supported databases is not negligible, economists are working on models that can demonstrate that the expense delivers a positive societal return.

For example, scientists at the U.S. Geological Survey (USGS) are developing an economic framework to measure what they call the "VOI" or value of information contained in the remarkable storehouse of Land Use / Land Cover maps created from Landsat's moderate resolution land imagery (or MRLI) since the early 1970s. In a test project involving 35 counties in eastern Iowa, scientists are using satellite imagery from the Landsat archive to observe historical crop rotation patterns of local agricultural lands. This land use history is linked to groundwater quality data from wells maintained by the Iowa Department of Natural Resources and the USGS to estimate potential changes to chemical concentrations in groundwater supplies. Due in part to biofuels initiatives and demands, there has been a recent increase in corn production. However, because growing corn demands higher nitrogen fertilizer usage than many other crops, there is a consequence for water quality as nitrates seep into the groundwater [(36)].

Researchers are using MRLI to develop forecasting models that will allow them to estimate the expected impact on water quality of current and recent planting decisions. The satellite imagery is particularly useful because the type of water contamination at issue here is so-called "nonpoint" source pollution, which is the result of the accumulation of the consequences of many different actors over a long period of time, as opposed to pollution occurring as a consequence of a single actor or event, such as a sewage treatment plant's output, for example.

Ultimately, such forecasts will assist policymakers in determining if and when it is cost effective to intervene in land use decisions that could affect the availability of the water resource. By reconciling groundwater pollution hazards with the region's agricultural needs, the USGS is aiming for a VOI calculation that can inform decisions that maximize agricultural production as well as lower mitigation and treatment costs necessary to avoid human health and other consequences of contaminated groundwater. In several smaller studies, USGS found the socioeconomic benefits provided by geologic maps were fully cost justifiable. Geologic maps have been shown to offer a positive VOI in other land use decisions as well, such as where to locate a county landfill, and a freeway retaining wall, as well as where to target mineral exploration.

Conclusion

The Task Force has evaluated the need for the NSF to create a sustainable data infrastructure to support world-class research and innovation and grouped the requirements for such an infrastructure under six broad areas of concern:

(1) Infrastructure Delivery;

(2) Culture and Sociological Change;

(3) Roles and Responsibilities;

(4) Economic Value and Sustainability;

(5) Data Management guidelines;

(6) Privacy, Confidentiality and Intellectual Property.

The key recommendations in each of these areas are summarized in the Executive Summary.

The Task Force believes that such a data infrastructure is essential to sustain the US's long-term leadership in scientific research and will provide a platform that will drive future discoveries, innovation and, ultimately, U.S. national prosperity. In addition, the creation and sustained funding of this infrastructure as a "first class citizen" in an integrated way rather than, as now, as an *ad hoc* collection of largely disconnected small-scale and short-term initiatives, will bring many other benefits. These benefits include improved repeatability and reproducibility of science; increased access and support for the medium-scale data part of the long tail of the scientific research community; and enhanced potential for citizen science engagement and participation. The NSF must also develop an educational pipeline to create a workforce with world-class skills in large-scale data management, curation, and analysis since these will constitute a critical skill base for both scientific research and for industry. The training programs that NSF establishes around such a data infrastructure initiative will create a new generation of data scientists, data curators, and data archivists that is equipped to meet the challenges and jobs of the future.

To again emphasize some of the key findings of the Task Force:

1. Social and cultural changes are as important -- if not more important -- than new technology. New incentive mechanisms will be essential.
2. Collecting more and more data without developing tools for analysis does not advance science. When we increase rewards for data contributions, it is important to also include rewards for building infrastructure and tools. Visualization is one important area of analytical capacity that we need to invest in.
3. There are variations across disciplines and areas of science, but we should start by satisfying common needs that we can extract from workshop reports and other sources.
4. Research libraries, archives, and other stewardship institutions have the capacity to aggregate and hold data, manage metadata, deal with rights management and access, and help users. This capacity needs to be modernized and fine-tuned for scientific data but should not be overlooked or replaced.

The Task Force believes that its key recommendations are both specific and pragmatic and will, if implemented by NSF, lead to a significant acceleration in the progress of science. Indeed, they stand to increase the research community's ability to find solutions for the problems facing the U.S. and, indeed, the whole of humanity. It is critical to keep in mind one important caveat, however: In order to reap the full benefit of this new data infrastructure there is a need for cultural change in the research community, in universities, in libraries, and also in funding agencies. The NSF needs to lead by example and catalyze such cultural change by creating an incentive system that rewards open and inclusive behavior and encourages the adoption of a more long-term perspective by both the research community and funders.

Acknowledgments

The authors are grateful for assistance and support from Becky O'Brien, Bailey Oland, and JoAnne Kerr at the University of Michigan and Drs. Daron Green, Jim Pinkelman and Evelyne Viegas at Microsoft Research.

Works Cited

1. Edwards, P. N., et al., et al. Report of a Workshop on “History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures”. *Understanding Infrastructure: Dynamics, Tensions, and Designs.* January 2007. http://epl.scu.edu/~gbowker/cyberinfrastructure.pdf.

2. NSF 07-28, Cyberinfrastructure Vision for 21st Century Discovery. *National Science Foundation Publications.* March 28, 2007. http://www.nsf.gov/pubs/2007/nsf0728/index.jsp.

3. NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. *National Science Foundation Publications.* September 2005. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.

4. Fostering Learning in the Networked World: The Cyberlearning Opportunity and Challenge. *National Science Foundation Publications.* August 11, 2008. http://www.nsf.gov/pubs/2008/nsf08204/nsf08204.pdf.

5. Interagency Working Group on Digital Data to the National Science and Technology Council. Harnessing the Power of Digital Data for Science and Society. *NITRD.* January 2009. http://www.nitrd.gov/About/Harnessing_Power_Web.pdf.

6. Force, Blue Ribbon Task. Ensuring Long-Term Access to Digital Information. *BRTF Web Site.* February 2010. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

7. Workshop, NSF. Data-Enabled Science in the Mathematical and Physical Sciences. *CRA Publications.* March 29-30, 2010. http://www.cra.org/ccc/docs/reports/DES-report_final.pdf.

8. Ahrens, Jim, et al., et al. *Data Intensive Science in the Department of Energy.* Los Alamos : Los Alamos National Laboratory, 2010. LA-UR-10-07088.

9. Hey, Tony, Tansley, Stewart and Tolle, Kristin, [ed.]. *The Fourth Paradigm: Data Intensive Scientific Discovery.* Microsoft Research, 2009. 978-0982544204.

10. Riding the Wave: How Europe can gain from the rising tide of scientific data. *Europe's Information Society Thematic Portal.* October 2010. http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707.

11. *IRIS.* 2010. http://www.iris.edu/hq/.

12. *Inter-University Consortium for Political and Social Research.* University of Michigan, 2010. http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp.

13. SageCite Project. http://blogs.ukoln.ac.uk/sagecite/.

14. Galaxy Zoo. http://www.galaxyzoo.org/.

15. World Wide Telescope. Microsoft Research, 2010. http://www.worldwidetelescope.org.

16. Google.org Flu Trends. Google, 2009. [Cited: Decemeber 5, 2010.] http://www.google.org/flutrends/.

17. Many Eyes. IBM. http://www-958.ibm.com/software/data/cognos/manyeyes/.

18. Lyon, Liz. UKOLN. *Dealing with Data: Roles, Rights, Responsibilities and Relationships.* June 19, 2007.

19. UKOLN. *UK Data Management Plan Online Tool.* UK Data Management Centre, 2010. http://dmponline.hatii.arts.gla.ac.uk/.

20. Beagrie, Neil, Lavoie, Brian and Woollard, Matthew. Keeping Research Data Safe 2. *JISC web site.* April 2010. http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf.

21. *Creative Commons.* http://creativecommons.org/licenses/by/2.5/.

22. Murray-Rust, Peter. *Panton Principles.* July 2009. http://pantonprinciples.org/about/.

23. *United States National Census Data.* 2010. http://www.census.gov/.

24. Guide to Open Data Licensing. Open Definition, 2010. http://www.opendefinition.org/guide/data/.

25. Kawamoto, Dawn and Mills, Elinor. CNET News. *CNET.* CNET, August 7, 2006. http://news.cnet.com/AOL-apologizes-for-release-of-user-search-data/2100-1030_3-6102793.html?tag=nefd.top.

26. What Information is "Personally Identifiable"? Electronic Frontier Foundation, 2010. http://www.eff.org/deeplinks/2009/09/what-information-personally-identifiable.

27. Wagner, Erika L and Kupriyanova, O. Data-driven Ethics: Exploring Customer Privacy in the Information Era. *The Center for Hospitality Research.* 2007. http://www.hotelschool.cornell.edu/research/chr/pubs/reports/abstract-14484.html.

28. Rampadarath, H., et al., et al. Hanny's Voorwerp: Evidence of AGN activity and a nuclear starburst in the central regions of IC 2497. *Astronomy & Astrophysics.* June 21, 2010, arXiv:1006.4096v1 [astro-ph.GA] 21 Jun 2010.

29. Dean, Cornelia. Scientific Savvy? In U.S., Not Much. *New York Times - Science.* New York Times, August 30, 2005. Scientific Savvy? In U.S., Not Much.

30. *About OldWeather.* 2010. http://www.oldweather.org/.

31. *Nearby Supernova.* http://snfactory.lbl.gov/.

32. *Alzheimer's Disease Neuroimaging Initiative.* http://www.adni-info.org/.

33. *Large Hadron Collider.* http://lhc.web.cern.ch/lhc/.

34. *Sunfall.* http://snfactory.lbl.gov/snf/snf-sunfall.html.

35. *Digital Curation Centre.* http://www.dcc.ac.uk/.

36. Bernknopf, R., W. Forney, R. Raunikar, and S. Mishra, 2011, A general framework for estimating the benefits of Moderate Resolution Imagery in environmental applications, in R. Laxminarayan and M. Macauley (ed.), Value of Information: Frontiers and New Applications, *Resources for the Future*, Washington, DC, in press.

Broader Contextual Bibliography

Overview

This selective bibliography organizes literature into five categories: Cyberinfrastructure, Data Curation/Stewardship, Digital Preservation, Sustainability and a category for items of a more general nature. The bibliography emphasizes recent literature while largely omitting contributions with a purely technical focus.

General

Anderson, C. (2008, June). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired, 16*(7). Available at: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Barksdale, J., & Berman, F. (2007, May 16). Saving our Digital Heritage. *The Washington Post*. Available at: http://www.washingtonpost.com/wp-dyn/content/article/2007/05/15/AR2007051501873.html

Barkstrom, B., & Sidell, P. (2009). *Accounting for the Value of Earth Science Data.* Available at: http://www.sciops.esa.int/SYS/CONFERENCE/include/pv2009/papers/28_Barkstrom_Accounting_For_Value.pdf

Bell, G., Hey, T., & Szalay, A. (2009, March 6). Beyond the Data Deluge. *Science, 323*, pp. 1297-1298. Available at: http://www.sciencemag.org/cgi/content/full/323/5919/1297

Bohn, R., & Short, J. (2009). *How Much Information? 2009: Report on American Consumers.* Available at: http://hmi.ucsd.edu/pdf/HMI_2009_ConsumerReport_Dec9_2009.pdf

Bollier, D. (2010). *The Promise and Peril of Big Data.* The Aspen Institute. Available at: available at: http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/InfoTech09.pdf

Council of the the European Union: 2832nd Competitiveness Council. (2007). *Council Conclusion on Scientific Information in the Digital Age.* Available at: http://www.consilium.europa.eu/uedocs/cms_Data/docs/pressdata/en/intm/97236.pdf

Dean, C. (2005) *Scientific Savvy? In U.S., Not Much*. New York Times - Science. New York Times, August 30, 2005.

Gray, J., & Szalay, A. (2007). *eScience - A Transformed Method.* Computer Science and Technology Board of the National Research Council. Available at: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm; Data-Intensive Scientific Discovery.* Redmond, Washington: Microsoft Research. Available at: http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Hey, T., & Trefethen, A. (2003). The data deluge: an e-science perspective. In F. Berman, G. Fox, & T. Hey, *Grid computing: Making the global infrastructure a reality* (pp. 809-924). Wiley and Sons. Available at: http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf

Hey, T. (2010). *The Next Scientific Revolution*. *Harvard Business Review*.

Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. (2009). *Harnessing the Power of Digital Data for Science and Society.* Available at: http://www.rin.ac.uk/system/files/attachments/Patterns_information_use-REPORT_Nov09.pdf

Jaschik, S. (2009). Digital Archives That Disappear. *Inside Higher Ed*. Available at: http://www.insidehighered.com/news/2009/04/22/record

JISC (2004). The Data Deluge: Preparing for the explosion in data. Available at: http://www.jisc.ac.uk/media/documents/publications/datadelugebp.pdf

JISC. (2007). *Dealing with Data: Roles, Rights, Responsibilities and Relationships.* Available at: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

National Science Board. (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century.* National Science Foundation. Available at: http://www.nsf.gov/pubs/2005/nsb0540

NSF Task Force on Cyberlearning. (2008). *Fostering Learning in the Networked World: The Cyberlearning Opportunity and Challenge.* Available at: http://www.nsf.gov/pubs/2008/nsf08204/nsf08204.pdf

Palm, J. (2006). *The Digital Black Hole,* available at: http://www.tape-online.net/docs/Palm_Black_Hole.pdf

Rampadarath, H., et al., (2010). *Hanny's Voorwerp: Evidence of AGN activity and a nuclear starburst in the central regions of IC 2497*. Astronomy & Astrophysics. June 21, 2010, arXiv:1006.4096v1 [astro-ph.GA] 21 Jun 2010.

Szalay, A., & Blakeley, J. (2009). Gray's Laws: Database Centric Computing in Science. In T. Hey, S. Tansley, & K. Tolle (Eds.), *Fourth Paradigm* (pp. 5-11). Redmond, Washington: Microsoft Research. Available at: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part1_szalay.pdf

The Economist (2010, February 27). Data, data everywhere. A special report on managing information. Available (with subscription) at: http://www.economist.com/node/15557443

The Economist (2010, February 27). All too much: Monstrous amounts of Data. Available (with subscription) at: http://www.economist.com/node/15557421

The Economist (2010, February 27). New rules for big data: Regulators are having to rethink their brief Available (with subscription) at: http://www.economist.com/node/15557487

Thomas, J. C. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics.* Available at: http://nvac.pnl.gov/docs/RD_Agenda_VisualAnalytics.pdf

Van de Sompel, H., & Lagoze, C. (2009). All aboard: Toward a machine-friendly scholarly communication system. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Redmond, Washington: Microsoft Research. Available at: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part4_sompel_lagoze.pdf

Wagner, E. L. and Kupriyanova, O. (2007). *Data-driven Ethics: Exploring Customer Privacy in the Information Era*. The Center for Hospitality Research. [Online] 2007. http://www.hotelschool.cornell.edu/research/chr/pubs/reports/abstract-14484.html.

Cyberinfrastructure

American Council on Learned Societies. (2006). Our Cultural Commonwealth. *ACLS Commission on Cyberinfrastructure.* Available at: http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf

Berman, F. (2008, July/August). Making Research and Education Cyberinfrastructure Real. *EDUCAUSE Review*. Available at: http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume43/MakingResearchandEducationCybe/163058

e-Infrastructure Reflection Group. (2009). *e-IRG white paper 2009.* Available at: http://www.e-irg.eu/publications/white-papers.html

First Monday (2007). Special Issue on Cyberinfrastructure, June 4, 2007, Volume 12 Number 6. Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/issue/view/240

Giaretta, A. (2009). *Components for a Science Data Infrastructure – preservation and re-use of data.* Available at: http://www.sciops.esa.int/SYS/CONFERENCE/include/pv2009/papers/31_Giaretta_ComponentsForScienceDataInfrastructure.pdf

Lynch, C. (2008). The Institutional Challenges of Cyberinfrastructure and E-Research. *EDUCAUSE Review*. Available at: http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume43/TheInstitutionalChallengesofCy/163264

Marek, K., Pires, C., & Glinos, K. (2009). *Scientific Data e-Infrastructures in the European Capacities Programme.* Available at: http://www.sciops.esa.int/SYS/CONFERENCE/include/pv2009/papers/48_Marek_EC-ScientificDataeInfrastructure.pdf

National Science Foundation - Cyberinfrastructure Council. (2007). *Cyberinfrastructure for the 21st Century Discovery.* Available at: http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

Office of Science and Innovation. (2007). *Developing the UK e-Infrastructure for Science and Innovation.* Available at: http://www.nesc.ac.uk/documents/OSI/report.pdf

Weidman, S., & Arrison, T. (2010). *Steps Toward Large-Scale Data Integration in the Sciences: Summary of a Workshop.* Washington, D.C.: National Academies Press.

Data Curation / Stewardship

Ailamaki, A., Kantere, V., & Dash, D. (2010, June). Managing Scientific Data. *Communications of the ACM*, 68-78. Available at: http://cacm.acm.org/magazines/2010/6/92486-managing-scientific-data/fulltext

Beagrie, N., Chruszcz, J., & Lavoie, B. (2008). *Keeping Research Data Safe (Phase 1).* Available at: http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

Beagrie, N., Lavoie, B., & Woollard, M. (2010). *Keeping Research Data Safe (Phase 2).* Available at: http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

COSEPUP (Committee on Science, Engineering, and Public Policy). (2009). *Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age.* Washington, DC: National Academies Press. Executive summary available at: http://www.nap.edu/html/12615/12615_EXS.pdf

Data Management Task Force. (2009). *e-IRG report on Data Management.* e-Infrastructure Reflection Group. Available at: http://www.e-irg.eu/images/stories/e-irg_dmtf_report_final.pdf

Friedlander, A., & Adler, P. (2006). *To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering.* Washington, D. C.: ARL. Available at: http://www.arl.org/bm~doc/digdatarpt.pdf

Gold, A. (2010). *Data curation and Libraries: Short-Term Developments, Long-Term Prospects.* Available at: http://works.bepress.com/agold01/9/

Gray, J., Szalay, S., Thakar, A., Stoughton, C., & Vandenberg, J. (2002). *Online Scientific Data Curation, Publication, and Archiving.* Available at: http://research.microsoft.com/pubs/64568/tr-2002-74.pdf

International Journal of Digital Curation (2006-Current). Available at: http://www.ijdc.net/index.php/ijdc

Key Perspectives. (2010). *Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study.* Digital Curation Centre. Available at: http://www.dcc.ac.uk/docs/publications/SCARP%20SYNTHESIS.pdf

Lindley, D. (2009, October). Managing Data. *Communications of the ACM*, 11-13. Available at: http://cacm.acm.org/magazines/2009/10/42492-managing-data/fulltext

Research Information Network. (2008). *Stewardship of digital research data: a framework of principles and guidelines. Responsibilities of research institutions and funders, data managers, learned societies and publishers.* London. Available at: http://www.rin.ac.uk/files/Research%20Data%20Principles%20and%20Guidelines%20full%20version%20-%20final.pdf

Research Information Network. (2008). *To Share or not to Share - RIN study on publication and quality assurance of research data outputs.* Available at: http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf

Research Information Network and British Library. (2009). *Patterns of information use and exchange: case studies of researchers in the life sciences.* Available at: http://www.rin.ac.uk/system/files/attachments/Patterns_information_use-REPORT_Nov09.pdf

Treloar, A., Groenewegen, D., & Harboe-Ree, C. (2007). The Data Curation Continuum: managing data objects in institutional repositories. *Dlib*. Available at: http://www.dlib.org/dlib/september07/treloar/09treloar.html

Witt, M. (2010). Institutional Repositories and Research Data Curation in a Distributed Environment. *Library Trends*, 191-201. Available at: http://muse.jhu.edu/journals/library_trends/v057/57.2.witt.pdf

Digital Preservation

Abrahamson, M., Bollen, K., Gutmann, M., King, G., & & Pienta, A. (2009). Preserving Quantitative Research-Elicited Data for Longitudinal Analysis. New Developments in Archiving Survey Data in the U.S. *Historical Social Research, 34*(3), 51-59.

Albani, S., & Giaretta, D. (2009). *Long Term Data and Knowledge preservation to guarantee access and use of the Earth Science archive.* Available at: http://www.sciops.esa.int/SYS/CONFERENCE/include/pv2009/papers/5_Albani_LongTermDataAndKnowledgePreservationForEarthScience.pdf

Altman, M. (2009). Transformative Effects of NDIIPP: the Case of the Henry A. Murray Archive. 338-351. Available at: http://muse.jhu.edu/journals/library_trends/v057/57.3.altman.html

Altman, M., Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., et al. (2009). Digital Preservation Through Archival Collaboration: The Data Preservation Alliance for the Social Sciences. *American Archivist*, 169-182. Available at: http://maltman.hmdc.harvard.edu/papers/SharedPractices.pdf

Baker, M., Shah, M., Rosenthal, D., Roussopoulos, M., Maniatis, P., Giuli, T., et al. (2006). *A Fresh Look at the Reliability of Long-Term Digital Storage.* Available at: http://www.nesc.ac.uk/talks/763/roussopoulos-presdb07.pdf

Berman, F. (2008). Got Data? A Guide to Data Preservation in the Information Age. *Communications of the ACM*, 50-58. Available at: http://www.nesc.ac.uk/talks/763/roussopoulos-presdb07.pdf

Chapman, S. (2003). Counting the Costs of Digital Preservation: Is Repository Storage Affordable. *Journal of Digital Information*. Available at: http://jodi.tamu.edu/Articles/v04/i02/Chapman/chapman-final.pdf

Day, M. (2008). Toward Distributed Infrastructures for Digital Preservation: The Roles of Collaboration and Trust. *The International Journal of Digital Curation*. Available at: http://www.ijdc.net/index.php/ijdc/article/view/60/39

Digital Preservation Coalition. (2010). *Digital Preservation - Preservation Issues.* Available at: http://www.dpconline.org/advice/digital-preservation-preservation-issues.html

Digital Preservation Coalition. (2010). *Digital Preservation Handbook: Strategic Issues.* Available at: http://www.dpconline.org/graphics/digpres/stratoverview.html

Gutmann, M., Abrahamson, M., Adams, M., Altman, M., Arms, C., Bollen, K., et al. (2009). From Preserving the Past to Preserving the Future: The Data-PASS Project and the Challenges of Preserving Digital Social Science Data. *Library Trends, 57*(3), 315-337. Available at: http://muse.jhu.edu/journals/library_trends/v057/57.3.gutmann.pdf

Heery, R., & Powell, A. (2006). *Digital Repositories Roadmap: looking forward.* Available at: http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/rep-roadmap-v15.pdf

Hoorens, S., Rothenberg, J., van Oranje, C., van der Mandele, M., & Levitt, R. (2007). *Addressing the uncertain future of preserving the past: Towards a robust strategy for digital archiving and preservation.* International Conference on Digital Preservation Tools and Trends. Available at: http://www.rand.org/pubs/technical_reports/TR510/

Lavoie, B., & Dempsey, L. (2004). Thirteen Ways of Looking at...Digital Preservation. *D-Lib Magazine*. Available at: http://www.dlib.org/dlib/july04/lavoie/07lavoie.html

Pirani, J., & Spicer, D. (2010). *The Chronopolis Project: A Grid-Based Archival Digital Preservation Solution (Case Study 1, 2010).* EDUCAUSE. Boulder, CO: EDUCAUSE Center for Applied Research. Available at: http://www.educause.edu/ir/library/pdf/ECAR_SO/ecs/dataman/ECS1001.pdf

Rosenthal, D. (2009). *How Are We Ensuring the Longevity of Digital Documents.* Available at: http://brtf.sdsc.edu/biblio/CNI2009plenary.pdf

SNIA Data Management Team. (2007). *100 Year Archive Task Force: Overview.* Storage Networking Industry Association (SNIA). Available at: http://www.snia.org/forums/dmf/programs/ltacsi/100_year/100-Yr-Archive-Task-Force-Overview_20060927.pdf

Sustainability

Arms, C. R., & Fleischauer, C. (2005). *Digital Formats: Factors for Sustainability, Functionality, and Quality.* IS & T Archiving 2005 Conference, Washington, D. C. Available at: http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf

Beddoe, R., Costanza, R., Farley, J., Garza, E., Kent, J., Kubiszewski, I., et al. (2009). *Overcoming systemic roadblocks to sustainability: The evolutionary redesign of worldviews, institutions, and* *technologies.* PNAS. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2650289/pdf/zpq2483.pdf

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2008). *Sustainable Economics for a Digital Planet: Issues and Challenges of Economically Sustainable Digital Preservation.* Interim Report. Available at: http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010). *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information.* Final Report. Available at: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Bradley, K. (2007). Defining Digital Sustainability. *Library Trends* , 148-163. Available at: http://muse.jhu.edu/journals/library_trends/v056/56.1bradley.pdf

Choudhury, S., & Hanisch, R. (2009). *The Data Conservancy: Building a Sustainable System for Interdisciplinary Scientific Data Curation and Preservation.* Available at: http://www.sciops.esa.int/SYS/CONFERENCE/include/pv2009/papers/47_Choudhury_DataConservancy.pdf

Eakin, L., Friedlander, A., Schonfeld, R., & Choudhury, S. (2008). *A Selective Literature Review on Digital Preservation Sustainability.* Available at: http://brtf.sdsc.edu/biblio/Cost_Literature_Review.pdf

Guthrie, K., Griffiths, R., & Maron, R. (2008). *Sustainability and Revenue Models for Online Academic Resources.* Available at: http://www.ithaka.org/ithaka-s-r/strategy/sca_ithaka_sustainability_report-final.pdf

Houghton, J., Rasmussen, B., Sheehan, P., Victoria University, Oppenheim, C., Morris, A., et al. (2009). *Economic Implications of Alternative Scholarly Publishing Models: Exploring the Costs and Benefits.* Available at: http://www.jisc.ac.uk/media/documents/publications/rpteconomicoapublishing.pdf

Hunter, L. (2006). *Investment in an Intangible Asset.* Available at: http://www.dcc.ac.uk/resource/curation-manual/chapters/intangible-asset/

Lavoie, B. (2003). *The Incentives to Preserve Digital Materials: Roles, Scenarios, and Economic Decision-Making.* OCLC Research. Available at: http://www.oclc.org/research/activities/past/orprojects/digipres/incentives-dp.pdf

Lavoie, B. (2004). Of Mice and Memory: Economically Sustainable Preservation for the 21st Century. In *Access in the Future Tense.* Council on Library and Information Resources. Available at: http://www.clir.org/pubs/reports/pub126/lavoie.html

Lavoie, B. (2008). The Fifth Blackboard: Some Thoughts on Economically Sustainable Digital Preservation. *D-Lib Magazine*. Available at: http://www.dlib.org/dlib/march08/lavoie/03lavoie.html

Ng, Y., Rubin, N., & Van Malssen, K. (2010). Strategies for Sustainable Preservation of Born Digital Public Television. Available at: http://www.thirteen.org/ptvdigitalarchive/files/2009/10/PDPTV_SustainabilityStrategies.pdf

Task Force on Sustainable e-Infrastructures. (2006). *e-Infrastructure Reflection Group (e-IRG) Task Force on Sustainable e-Infrastructures (SeI).* e-Infrastructure Reflection Group. Available at: http://www.e-irg.eu/images/stories/reports/2006-report_e-irg_tf-sei.pdf

http://www.nsf.gov/od/oci/taskforces